Information Integration from Heterogeneous Data Sources: a Semantic Web Approach

Narendra Kunapareddy, Parsa Mirhaji, David Richards, S. Ward Casscells
Center for Biosecurity and Public Health Informatics Research, Houston, TX

Abstract
Although the decentralized and autonomous implementation of health information systems has made it possible to extend the reach of surveillance systems to a variety of contextually disparate domains, public health use of data from these systems is not primarily anticipated [1]. The Semantic Web has been proposed to address both representational and semantic heterogeneity in distributed and collaborative environments [2]. We introduce a semantic approach for the integration of health data using the Resource Definition Framework (RDF) and the Simple Knowledge Organization System (SKOS) [3] developed by the Semantic Web community.

Introduction
The Semantic Web is a framework specifically designed to foster information sharing and multidisciplinary (re)use of informational resources in collaborative and distributed environments such as the World Wide Web. The Resource Definition Framework (RDF) provides a general purpose framework for representing information, with a schema language, RDF(S), that provides the basic ingredients for a shared representation.
The Simple Knowledge Organization System (SKOS) is a set of specifications and standards to support the use of knowledge organization systems such as thesauri, classification schemes, subject heading lists, taxonomies and other types of controlled vocabulary, and possibly terminologies and glossaries, all within the framework of the Semantic Web.

Methods
Overview: Our method uses the XML structure of incoming data and a reference ontology to automatically derive a conceptual graph representing the semantics of the data. For each data source, a conceptual graph is created according to the SKOS framework. This graph, which we call a "Simple Concept Organization System" (SCOS), is mapped (semi-automatically) to a set of integration ontologies for semantic integration. Integration ontologies can be extended to repurpose or share data and to support different tasks across disparate applications.
Schematization: An automated agent evaluates incoming XML messages and compares information element by element and attribute by attribute against the dynamically-built SCOS using a reference ontology. The reference ontology is implemented in OWL-DL and contains a model of XML document structure, and rules and axioms to interpret a generic XML document.
New concepts detected in XML documents are indexed using the SKOS:broader, SKOS:narrower, SKOS:related, and SKOS:definition relationships between SKOS:Concept(s). The schematizer ensures that every concept within incoming data has a unique representation in SCOS.
Integration: Each data instance is stored as an instance of a SCOS Concept, each of which has relations to other SCOS Concepts that represent metadata such as time stamps and source data.

Results
An integrated RDF repository has been built from data submitted by eight community hospitals, consisting of all triage and medical records entries from emergency department visits; a combination of structured, semi-structured, and non-structured data. All changes in the schema or vocabularies used in the source are captured automatically and reflected in the SCOS repository. The SCOS representation enables use of integration ontologies to support multidisciplinary use of the data, making the original context of the data computationally available for query planning and execution.
The integrated SCOS repository lends itself to the application of any kind of rule-based or ontology-based queries, and subgraph and link analysis algorithms using standard RDF query languages.

References
1. Sujansky W. Heterogeneous Database Integration in Biomedicine. Journal of Biomedical Informatics. 2001;34,:285–98.
2. Vdovjak R, Eindhoven G-JH. RDF Based Architecture for Semantic Integration of Heterogeneous Information Sources. Workshop on Information Integration on the Web; 2001; Eindhoven University of Technology; 2001.
3. Miles A, Brickley D. Simple Knowledge Organisation System (SKOS). Nov 2005 [cited; Available from: http://www.w3.org/TR/2005/WD-swbp-skos-core-guide-20051102